

ДИСПЕРСІЙНИЙ АНАЛІЗ

Змістовий модуль 4

Основні поняття

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії. У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англomовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

Факторами називають контрольовані чинники, що впливають на кінцевий результат. **Рівнем фактора**, або способом обробки, називають значення, що характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають **відгуком**.

Часто вихідні значення факторів вимірюють у **кількісних** або **порядкових** шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактора.

Якщо кількість груп взяти надмірно великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів. Якщо її взяти надмірно малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему. Вибір конкретного способу групування даних залежить від їх обсягу і характеру варіювання значень фактора.

Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот. При багатофакторному аналізі застосовують три типи групування:

- ❑ **групи з рівною кількістю спостережень;**
- ❑ **групи з різною кількістю спостережень;**
- ❑ **групи, кількості спостережень у яких відповідають певній пропорції.**

Однофакторний аналіз

Основною метою **однофакторного аналізу** зазвичай є оцінка величини впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають **контрастом факторів**. Попереднім етапом є перевірка нульової гіпотези **про відсутність будь-якого впливу досліджуваного фактора (факторів)**, тобто гіпотези про те, **що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності**.

Якщо нульову гіпотезу відкидають, то наступним етапом є **кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик**. У випадку, коли нульова гіпотеза не може бути відкинута, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може впливати з теоретичних уявлень про об'єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

При однофакторному дисперсійному аналізі вихідні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень при відповідному рівні фактора. Для різних рівнів фактора кількість спостережень може бути різною. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є **перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються**.

Форма таблиці спостережень при проведенні однофакторного дисперсійного аналізу

Результати вимірювань	Рівні фактора			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
			...	
n_i	x_{ni1}	x_{ni2}	...	x_{nik}

Вплив дієти (фактор) на зміну маси тіла (відгук)

	A	B	C	D	E	F	G	H	I	J
1	Изменение веса									
2	Диета 1	Диета	Диета	Дисперсионный анализ: однофакторный						
3	3,8	0	7							
4	6	0	5,6	ИТОГО						
5	0,7	-2,1	3,4	Группы	Количество	Сумма	Среднее	Дисперсия		
6	2,9	2	6,8	Диета 1	24	79,2	3,300	5,018		
7	2,8	1,7	7,8	Диета 2	27	81,7	3,026	6,367		
8	2	4,3	5,4	Диета 3	27	139	5,148	5,739		
9	2	7	6,8							
10	8,5	0,6	7,2							
11	1,9	2,7	7	Дисперсионный анализ						
12	3,1	3,6	7,3	Источник дисперсии	Сумма квадратов	степень свободы	Квадрат среднего	F	Значение P	F критическое
13	1,5	3	0,9	Между группами	71,094	2,000	35,547	6,197	0,003	3,119
14	3	2	7,6	В группах	430,179	75,000	5,736			
15	3,6	4,2	4,1	Всего	501,273	77,000				
16	0,9	4,7	6,3							
17	-0,6	3,3	5							
18	1,1	-0,5	2,5							
19	4,5	4,2	0,9							
20	4,1	2,4	3,5							
21	9	5,8	0,5							
22	2,4	3,5	2,8							
23	3,9	5,3	8,6							
24	3,5	1,7	4,5							
25	5,1	5,4	2,8							
26	3,5	6,1	4,1							
27		7,9	5,3							
28		-1,4	9,2							
29		4,3	6,1							
30	3,3	3,0	5,1							
31	2,24	2,56	2,44							

Однофакторний аналіз

Непараметричним аналогом однофакторного дисперсійного аналізу є ранговий однофакторний аналіз **Краскела – Уолліса**. Він розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р. Цей критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві. Нульова гіпотеза полягає в тому, що k вибірок обсягами n_1, n_2, \dots, n_k отримані з однієї і тієї самої генеральної сукупності. Критерій Краскела – Уолліса є узагальненням U -критерію Манна – Уїтні на випадок, коли кількість вибірок $k > 2$.

Рангові методи, у тому числі й метод Краскела – Уолліса, **не передбачають нормальності розподілу** результатів спостережень і **можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак.**

Критерій Джонкхієра (Джонкхієра – Терпстра) запропонований незалежно один від одного нідерландським математиком Т.Дж. Терпстрою в 1952 р. й британським психологом Е.Р. Джонкхієром в 1954 р. Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактора, який вимірюють у порядковій шкалі. Таблиця даних має такий самий вигляд, як і в попередньому випадку. Будемо вважати, що її перший стовпчик відповідає найменшому рівню фактора, другий – наступному за величиною тощо, останній стовпчик відповідає найбільшому рівню. При виконанні таких припущень критерій Джонкхієра є більш потужним, ніж критерій Краскела – Уолліса, стосовно гіпотези про монотонний вплив фактора.

Однофакторний аналіз

М-критерій Бартлетта запропонований британським статистиком Маурісом Стівенсоном Бартлеттом в 1937 р. Його застосовують для перевірки нульової гіпотези про рівність дисперсій кількох нормальних генеральних сукупностей, з яких взяті досліджувані вибірки, що у загальному випадку мають різні обсяги (обсяг кожної вибірки має бути не менше чотирьох).

Г-критерій Кокрена (Кочрена) запропонований американським статистиком Вільмом Геммелом Кочреном в 1941 р. Його використовують для перевірки нульової гіпотези про рівність дисперсій k ($k \geq 2$) нормальних генеральних сукупностей за незалежними вибірками рівного обсягу.

Непараметричний **критерій Левене**, запропонований американським математиком Ховардом Левене в 1960 р. є альтернативою критерію Бартлетта в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковуються нормальному розподілу.

Двофакторний аналіз

Двофакторний дисперсійний аналіз застосовують для пов'язаних нормально розподілених вибірок. Дані подають у вигляді таблиці, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в рядках – дані, що відповідають рівням другого. Таблиця даних має розмірність $n \times k$, де n і k – кількість рівнів першого та другого факторів, відповідно.

Рівні фактора А	Рівні фактора В			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
			...	
n_i	x_{ni1}	x_{ni2}	...	x_{nik}

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив другого фактора є суттєвим. На практиці часто використовують і складніші таблиці двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

Двофакторний аналіз

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, то використовують непараметричний **ранговий критерій Фрідмана (Фрідмана, Кендалла та Сміта)**, розроблений американським економістом Мілтоном Фрідманом наприкінці 1930 р.

Q-критерій Кокрена запропонований В. Кочреном в 1937 р. Його використовують у випадках, коли групи однорідних суб'єктів піддаються впливам, кількість яких перевищує два, і для яких можливі два варіанти відгуків – умовно-негативний (0) та умовно-позитивний (1). Нульова гіпотеза полягає в рівності ефектів впливу.